**ОТЧЕТ**

**по хакатону «Осенней школы по информационным**

**технологиям ОИЯИ» на тему**

**«Аналитика больших данных: «физика высоких энергий»»**

Участники 13 группы Рылова А.И., Сулимов А.С., Трофимов И.А., Елистратов В.Е.

**Дубна**

**2024 г.**

**Содержание**

1. Цель работы и инструменты

2. Методология исследования

3. Результаты и выводы

4. Заключение

**1. Цель работы и инструменты**

**1.1 Цель работы**

Разработать программное решение для автоматизированного анализа предоставленного файла с публикационной активностью Китайской Народной Республики в области физики высоких энергий.

1.2 **Инструменты**

Язык — python

Библиотеки:

* numpy
* pandas
* pathlib
* json
* csv
* os

**2. Методология исследования**

**2.1 Описание переменных**

|  |  |  |
| --- | --- | --- |
| Тип | Имя | Описание |
| set | high\_energy\_physics\_terms | Множество, содержащее ключевые слова на тему «физика высоких энергий» |
| DataFrame | top\_organizations | Cловарь «топ 5 организаций по количеству статей» |
| str | main\_dir | Путь до папки с csv файлами |
| dict | organizations\_dict | Cловаpь вида: {Название института: как часто встречается} |

**2.2 Описание функций**

|  |  |  |
| --- | --- | --- |
| Имя функции | Входные данные | Описание функции |
| find\_articles | Множество ключевых слов: датафрейм, содержащий статьи | Ищет статьи, содержащие не менее двух ключевых слов |
| process\_file | Путь к файлу, словарь organizations\_dict | Создание словаря вида: {Название института: как часто встречается} |
| top\_5\_pairs | Словарь | Возвращает словарь «топ 5 организаций по количеству статей» |

**2. 3 Основные этапы работы программы:**

1. **Загрузка ключевых слов:** Создается список ключевых слов, характерных для физики высоких энергий.
2. **Чтение CSV-файлов:** Программа последовательно читает каждый CSV-файл в указанной директории.
3. **Фильтрация публикаций:** Для каждой публикации проверяется наличие ключевых слов.
4. **Извлечение и подсчет организаций:** Извлекаются названия организаций из отобранных публикаций и ведется их подсчет.
5. **Ранжирование:** Организации ранжируются по количеству упоминаний.
6. **Вывод результатов:** Выводятся топ-5 организаций с наибольшим количеством упоминаний.

2.4 **Методы исследования:**

1. **Сбор и подготовка данных:**
   * **Источник данных:** CSV-файлы, содержащие информацию о публикациях.
   * **Предобработка:**
     + Очистка данных от лишних столбцов и пропусков.
     + Приведение названий столбцов к единому формату.
     + Нормализация данных (например, приведение названий организаций к нижнему регистру).
2. **Фильтрация публикаций по тематике:**
   * **Ключевые слова:** Используется предопределенный список ключевых терминов, характерных для физики высоких энергий.
   * **Поиск совпадений:** В каждой публикации ищется наличие хотя бы двух ключевых слов из списка.
3. **Анализ организаций:**
   * **Извлечение названий организаций:** Извлекаются названия организаций из каждой публикации.
   * **Подсчет упоминаний:** Для каждой организации ведется подсчет количества упоминаний.
   * **Ранжирование:** Организации ранжируются по убыванию количества упоминаний.

**3. Результаты и выводы**

**3.1 Основные результаты**

В ходе анализа были выделены следующие организации с наибольшим количеством публикаций в области физики высоких энергий:

1. Guangxi Key Laboratory of Low Carbon Energy Materials, School of Chemistry and Pharmaceutical Sciences, Guangxi Normal University, Guilin

• Количество публикаций: 33

2. Key Laboratory of Luminescence Analysis and Molecular Sensing (Southwest University), Ministry of Education, College of Chemistry and Chemical Engineering, Southwest University, Chongqing

• Количество публикаций: 31

3. Key Laboratory of Luminescence Analysis and Molecular Sensing (Southwest University), Ministry of Education, School of Chemistry and Chemical Engineering, Southwest University, Chongqing

• Количество публикаций: 29

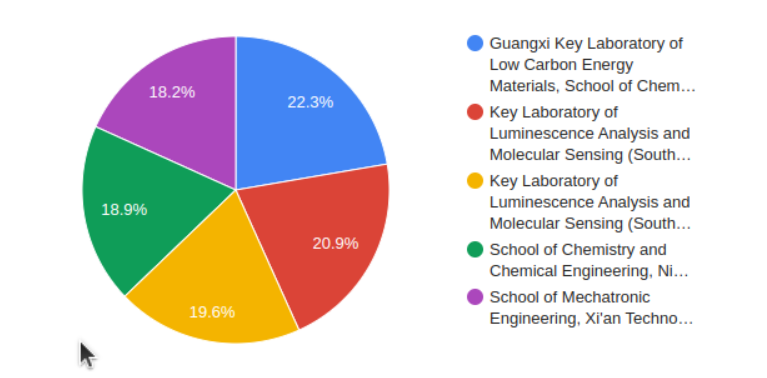
4. School of Chemistry and Chemical Engineering, Ningxia Key Laboratory of Solar Chemical Conversion Technology, Key Laboratory for Chemical Engineering and Technology, State Ethnic Affairs Commission, North Minzu University, Yinchuan

• Количество публикаций: 28

5. School of Mechatronic Engineering, Xi'an Technological University, Xi'an

• Количество публикаций: 27

Время выполнения программы: 4 минуты 45 секунд на M3 Pro 12 core

Рис. 1 «Диаграмма лидеров организаций по количеству статей»

**3.2 Анализ и выводы**

Лидеры публикационной активности:

• Наиболее активной организацией является Guangxi Key Laboratory of Low Carbon Energy Materials с 33 публикациями. Это может свидетельствовать о высоком уровне исследований и активном участии в международных научных проектах.

• Две лаборатории из Southwest University занимают второе и третье места по количеству публикаций, что указывает на значительное сотрудничество внутри данного университета.

**3.3 Код**

Код программы приложен на github.

<https://github.com/alice3e/JINR_autumn_school/blob/main/hacaton_2/hackaton_2_10_october.ipynb>

**Заключение**

Анализ публикационной активности организаций КНР в области физики высоких энергий выявил ключевых игроков и направления сотрудничества. Результаты могут служить основой для дальнейшего развития научных партнерств и междисциплинарных исследований.